ATIC FILE COPY

②

AIR FORCE

HUMAN RESOURCES

AD-A200 678

DTIC
ELECTE
NOV 2 1 1988
D ᵡᵦ

AIR FORCE OFFICER QUALIFYING TEST (AFOQT)
FORM P: TEST CONSTRUCTION

Frances R. Berger
Willa B. Gupta
Raymond M. Berger

Psychometrics Inc.
13245 Riverside Drive
Sherman Oaks, California 91423-2172

Jacobina Skinner

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

October 1988

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

88 11

NOTICE

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

HAROLD G. JENSEN, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE

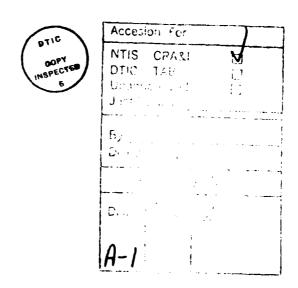Form Approved
OMB No 0704-0188

1a REPORT SECURITY CLASSIFICATION	1b RESTRICTIVE MARKINGS
Unclassified	

2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT
	Approved for public release;
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	distribution is unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)
	AFHRL-TR-88-30

6a NAME OF PERFORMING ORGANIZATION	6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION
Psychometrics Inc.		Manpower and Personnel Division

6c ADDRESS (City, State, and ZIP Code)	7b ADDRESS (City, State, and ZIP Code)
13245 Riverside Drive	Air Force Human Resouces Laboratory
Sherman Oaks, California 91423-2172	Brooks Air Force Base, TX 78235-5601

8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
Air Force Human Resouces Laboratory	HQ AFHRL	F33615-83-C-0035

8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
Brooks Air Force Base, Texas 78235-5601	62703F	7719	18	24

11. TITLE (Include Security Classification)

Air Force Officer Qualifying Test (AFOQT) Form P: Test Construction

12. PERSONAL AUTHOR(S)
Berger, F.R.; Gupta, W.B.; Berger, R.M.; Skinner, J.

13a. TYPE OF REPORT	13b TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Sep 83 TO May 86	October 1988	56

16 SUPPLEMENTARY NOTATION
Analyses were conducted in AFHRL/TS Study numbers S8703, S8858, S8864, S8871, S8907, S9057, S9063

17.	COSATI CODES		18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Air Force Officer Qualifying Test, officer selection,
05	08		aptitude tests, officer classification
05	09		mental abilities testing, test construction

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Since 1951, the Air Force Officer Qualifying Test (AFOQT) has been part of the selection process for officer commissioning programs and pilot and navigator training. The latest of 15 successive forms of the AFOQT, Form O, contains 380 items organized into 16 subtests which form five composites: Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative. The anticipated need for future forms of the AFOQT prompted the development of a large pool of experimental items. Approximately 4,800 items were developed and administered (along with Form O items) to basic airmen and officer cadets attending military training programs. The immediate goal was to construct a follow-on AFOQT Form P in two parallel forms. Item analyses of the experimental test results provided item difficulty and discrimination data that helped to guide the selection of new and Form O items for Form P. Resemblance to Form O in psychometric properties, content, and style was the criterion for Form P item selection. Preliminary analyses of the newly constructed forms (P1 and P2) indicated that they were highly similar to each other as well as being comparable to Form O in content, item difficulty, and item discriminative power. Follow-on research with additional military samples is recommended to assess more fully the parallelism among the three forms. (Keywords:)

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
Nancy J. Allin, Chief, STINFO Office	(512) 536-3977	AFHRL/TSR

DD Form 1473, JUN 86   Previous editions are obsolete   SECURITY CLASSIFICATION OF THIS PAGE

SUMMARY

This report describes the development of the Air Force Officer Qualifying Test (AFOQT), Form P. Since 1951, 15 successive forms of the AFOQT have been used to select individuals for officer commissioning programs and for pilot and navigator training. The latest AFOQT, Form O, was implemented in 1981 and consists of 16 subtests administered in a single test booklet. Percentile scores for five composites are derived from combinations of the subtests (Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative). In 1983, anticipation of the need for future forms of the AFOQT prompted a project to construct AFOQT Form P in two parallel forms. The motivation to develop these forms concurrently was to increase test security.

Approximately 4,800 items were created for an experimental pool in the initial stages of the project. The major requirement guiding development of new items for each of the 16 subtests was a high resemblance to Form O subtests in content, appearance, and distributions of item difficulty and discrimination. Items for each subtest were divided among seven experimental test booklets. Each booklet contained approximately 43 new items and 20 common items from Form O and was administered to samples of approximately 350 airmen attending Basic Military Training. Supplemental data for selected subtests were collected from samples of about 200 cadets at the Officer Training School. Item adequacy was evaluated using analyses of item difficulty, discrimination, and test reliability.

The selection of items for Form P was guided by item difficulty and discrimination requirements, as well as by the need to match Form O subtests for content and appearance. Preliminary analyses of the newly constructed forms (P1 and P2) revealed that they were highly similar. Distributions of items by difficulty, discrimination, and content were virtually identical. Further, the new forms were found to be comparable to Form O. Follow-on research is needed. Administration of the new forms to military samples is recommended to evaluate more fully the degree of parallelism between Form P1 and Form P2 and their statistical similarity to Form O.

PREFACE

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

LIST OF TABLES

AIR FORCE OFFICER QUALIFYING TEST (AFOQT)
FORM P: TEST CONSTRUCTION

I. INTRODUCTION

The officer candidate selection system for the U.S. Air Force involves several decision points, one of which is "mental qualification" as determined by scores on the Air Force Officer Qualifying Test (AFOQT). The test is used to select individuals for Officer Training School, Air Force Reserve Officer Training Corps cadets for scholarships or for the Professional Officers Course, and students for Undergraduate Pilot Training and Undergraduate Navigator Training. (Applicants to the Air Force Academy are exempt from this testing requirement.) Due to the importance of its use, the AFOQT requires periodic checks on its predictive validity, its currency, and its security, which in turn may determine the need for new forms of the battery. Anticipating this need, the Air Force Human Resources Laboratory (AFHRL) initiated a project to develop a large pool of new items from which new forms of the AFOQT could be assembled. This report describes the construction of two experimental parallel forms of the AFOQT, Forms P1 and P2. The Form P project was the first in the history of the AFOQT in which two parallel forms were to be constructed concurrently. The advantages of this approach are expected to include improved retesting capability and the reduction of opportunities for test compromise. The forms constructed in this project were experimental; follow-on research has been planned to assess their adequacy for use in the operational officer testing program as replacements for the single form (Form O) currently in use.

Psychological testing to select pilots for the U.S. Air Force was in use prior to World War II, but the actual forerunners of AFOQT Form P began in 1942 with the Aviation Cadet Qualifying Examination (ACQE) and the Aircrew Classification Battery (ACB), and in 1949 with the Aviation-Cadet Officer-Candidate Qualifying Test. The purpose of these tests was to help in the selection of commissioned officers and aircrew personnel. Modifications, combinations, and refinements of these tests followed annually at first, and then approximately every 3 years. Form A of the AFOQT, a multiple aptitude battery designed both to predict success in Officer Candidate School and to screen for aircrew training, was implemented in 1953 and was succeeded by 14 revised forms of the test. The history of the evolution of the AFOQT has been documented in several reports, the latest of which was prepared by Rogers, Roach, and Short (1986).

The development of a large pool of AFOQT items and the subsequent construction of Form P were aimed at emulating the style and content of Form O. The rationale for providing this continuity arises from the value of Form O for predicting success in training. Validity studies have demonstrated that Form O subtests and composites have correlated significantly with performance in non-rated technical training courses (Arth, 1985; Arth & Skinner, 1986; Finegold & Rogers, 1985) and in pilot and navigator training (Arth, Steuck, Sorrentino, & Burke, in preparation).

The AFOQT contains 16 subtests covering verbal, quantitative, perceptual, and specialized ability areas. Table 1 shows that approximately 50% of Form P was to consist of items drawn from Form O, a strategy designed to provide content continuity across forms. The items common to the two forms were to be augmented with the new experimental items to update and improve the quality of the Form P battery. Each of the parallel forms of Form P would be of the same length (380 items) and require the same total administration time (4.5 hours) as the current AFOQT Form O. Additional details about test content including sample items for each subtest have been published in the AFOQT Information Pamphlet (United States Air Force, 1987).

The next part of the report summarizes the development of the new experimental item pool. Part III indicates how items were selected for Form P, and Part IV provides recommendations with regard to the item pool and the parallel forms.

1

Table 1. Content and Organization of AFOQT Forms P1 and P2

Subtests	Form 0 common items	Number of new items	Total items P1, P2 (each)
Verbal Analogies	13	12	25
Arithmetic Reasoning	12	13	25
Reading Comprehension	13	12	25
Data Interpretation	12	13	25
Word Knowledge	13	12	25
Math Knowledge	12	13	25
Mechanical Comprehension	11	9	20
Electrical Maze	10	10	20
Scale Reading	20	20	40
Instrument Comprehension	10	10	20
Block Counting	10	10	20
Table Reading	20	20	40
Aviation Information	10	10	20
Rotated Blocks	7	8	15
General Science	10	10	20
Hidden Figures	8	7	15
Total	191	189	380

II. EXPERIMENTAL ITEM POOL

Method

Development of Experimental Tests

Item Writing. The development of the pool of new AFOQT items took into account the standard considerations for item writing, as explicated, for example, by Wesman (1971). Additional considerations that guided the appearance, position, and content of the new items were: consistency with Form 0 items in format, semantics, and punctuation; taxonomy of certain Form 0 subtests; item difficulty range; and specific requirements for graphics. Item writers, whether project staff or subject-matter consultants, were provided with operational definitions that described the constraints to be observed as to content, scope, complexity, length, appearance, and number of items and response options. When appropriate, the content categories covered in selected subtests were expanded to include a broader concept of the area tested; for example, adding computer-related questions to the General Science subtest because of their importance in the science fields. Taxonomies were developed, where appropriate, for the content of Form 0 tests so that writers of new items could duplicate the distribution of subject matter within content areas.

To create a pool of items that would result in the needed distribution of difficulty, the texts of Form 0 items, including stem and alternatives, and their item data[1] were analyzed to gain

---

[1] The sample on which test item data were based consisted of all first-time Form 0 examinees between 1 March 1982 and 29 February 1984 (N = 75,980).

further insight about factors that seemed to affect difficulty. These findings were then used to guide item construction to approximate the same range of difficulty as Form O within each subtest. A total of 4,800 experimental items were written, 300 for each subtest. The requirement for a large number of items was based on AFHRL experience that one in three items developed has the desired characteristics and meets the standards of acceptability for the AFOQT. These numbers ensured some latitude in identifying among the statistically acceptable items those that would most closely match a variety of Form O characteristics. As new items were constructed, they were pilot-tested informally, revised as necessary, and edited by the contractor's staff. Further critiquing and editing were accomplished by AFHRL and the items were returned to the contractor for assembly into test booklets.

Test Booklet Preparation. A total of 126 different test booklets was prepared. Seven booklets of items were assembled for each subtest except Scale Reading and Table Reading which were defined as speeded tests in Form O (Rogers, Roach, & Wegner, 1986). In order for item statistics for the last items of a speeded test to be as accurate as those for items that begin the test, it is important that all items be attempted. To ensure that this goal would be met, 14 booklets were prepared for these tests, seven with forward-order and seven with reverse-order item layouts. Each booklet in a subtest contained the same 20 items from the Form O subtest ("common" items) and 43 new items. The seven booklets of a subtest therefore contained 301 new items. Exceptions were the Rotated Blocks and Hidden Figures booklets, each of which consisted of 15 common and 43 new items; the Mechanical Comprehension booklets (19 common and 43 new); and the Block Counting booklets (20 common and 45 new).

Common Items. Common items, selected by AFHRL staff from Form O, provided a basis for verifying that the different samples tested on the seven booklets of a subtest were comparable in terms of ability levels. They also were useful for estimating the difficulty of new items for officer applicants. (These topics will be discussed in greater detail later.) The 20 common items followed the Form O order; that is, they were in the same location relative to each other in the experimental booklets as they were in the Form O test. The selection of Form O common items was based on data from the operational sample of 75,980 applicants for officer commissioning who took Form O between 1 March 1982 and 29 February 1984. Items with optimum ranges of difficulty and discrimination were the first to be considered. Format, content, and graphic characteristics were also important considerations but were secondary to the statistical criteria.

Item Try-Out Samples. Approximately 350 basic airmen were tested on each of the experimental booklets between August 1984 and December 1986. No subject took more than one booklet of a particular subtest. Several constraints precluded the use of a preferred sample (i.e., civilian applicants for Air Force officer commissions, the target population for which AFOQT Form P was designed). Since the AFOQT is administered for operational selection and classification purposes at about 500 military testing sites in the Continental United States and overseas, it was not logistically or economically feasible to try out the several thousand new test items with officer applicants. Basic airmen constituted the only practicable group on which to obtain preliminary data for evaluating item adequacy. The Basic Military Training (BMT) program has for many years provided a large and readily accessible source of examinees for AFHRL research and development (R&D) on skill and ability requirements for Air Force military occupations.

Supplemental data on items in two subtests -- General Science (GS) and Aviation Information (AI) -- were obtained by readministering the experimental test booklets to cadets attending Officer Training School (OTS) between October 1985 and January 1986. The GS and AI subtests assess knowledge in relatively technical and specialized areas. Results of the initial try-outs suggested that airmen found the test content to be quite difficult. They answered only about 29% to 36% of the items correctly in the various booklets. Item difficulty indices (proportion correct) fell below .30 for 41% to 65% of the items. These airmen performance levels prompted the establishment of special testing sessions with OTS cadets. Data obtained from the cadets were expected to

provide a sounder basis for evaluating the adequacy of new items in the GS and AI subtests. Cadets are baccalaureate degree holders and have typically completed 2 to 4 more years of formal education than the majority of airmen.

Administration Procedures

The collection of experimental item data was accomplished in test administration sessions lasting about 3 hours each. Multiple sessions were required to achieve the desired sample sizes for airmen and cadets. During each session it was typical for 45 airmen or 100 cadets to be tested on 2 to 3 booklets. Each booklet contained items from a different subtest; this procedure insured that the 7 (or 14) samples for each subtest were independent. Potential order-of-presentation effects were controlled by counterbalancing the sequence in which booklets were administered.

Time limits for each power subtest were determined after the first several administrations of any test by noting the number of minutes required for 95% of the examinees to finish that subtest. The average became the time limit for the subsequent administrations of the remaining booklets of that particular subtest. For the speeded subtests, the time limits were established based on the number of minutes required for 5% of the examinees to complete the test.

The practices and procedures used to administer Form O at operational test sites were observed as closely as possible during collection of experimental item data. Major features of the manual for administration for Form O were replicated; subtest directions were not changed, for example. Selected portions of the general instructions were modified, however, to enable collection of demographic information from examinees. These data were essential for a later evaluation of the comparability of the samples. Demographics and test responses were recorded on a machine-scannable answer sheet (General Answer Sheet Type C, Westinghouse Corporation, Form 093937-001 W-2300).

Analysis and Results

Sample Comparability

Demographic and aptitude characteristics of the airmen examinees were analyzed to assess the equivalence of the 7 (or 14) samples tested on each experimental subtest. Detailed results are reported in Appendix A. Comparability was essential to maximize the constancy of scale indices for item difficulty and discrimination measures across examinee groups. The samples, based on the collective set of findings described below, were judged to be sufficiently comparable to proceed with a Form P test construction strategy in which item indices were treated as non-sample-specific.

Age. The mean ages for the airmen samples ranged from 19.4 to 20.8 years (see Table A-1). Mean ages across samples for any subtest usually differed by no more than 4 to 10 months. Differences of 1 year to 14 months were found across the samples for three subtests (Electrical Maze, Block Counting, and Table Reading). The largest difference was for the General Science subtest (23 months).

Education. Most examinees (95%) had completed 12 to 14 years of formal education (see Table A-2). The mean number of years of education across the samples taking any subtest rarely varied by more than 2 months. The largest range of means was for the seven airmen samples who took the General Science subtest (12.2 to 12.7 years).

Gender. All of the AFOQT subtests and the majority of test booklets were administered to both males and females, but some booklets were administered only to males (see Table A-3). The range of group representation in the various samples for each subtest varied considerably for some subtests (e.g., Males, 55% to 100% for Arithmetic Reasoning) and very little for others (e.g., Males, 93% to 100% for General Science).

Ethnicity. The samples were racially mixed, with Whites forming the majority (75% to 80%). The minority representation was usually 10% to 15% Black, 3% to 5% Hispanic, 3% or less Asian or Pacific Islander, and 1% or less American Indian or Alaska native (see Table A-4). With respect to sample comparability, the range of differences in percentage representation was rarely more than 10%. In the Block Counting and Math Knowledge samples, White representation ranged from 73.5% to 84.5% and from 72.7% to 84.6%, respectively, but the typical pattern was for White representation to vary less than 9% across samples. Percentage differences for the other racial groups across samples were very small.

AFQT Scores. Comparability of the samples was also examined relative to performance on a measure used to screen for enlistment eligibilty, the Armed Forces Qualification Test (AFQT) composite score, which is derived from the Armed Services Vocational Aptitude Battery (ASVAB). Average AFQT percentile scores for the item try-out samples ranged from 59.1 to 67.3. Standard deviations were 16 to 19 points, indicating that the try-out samples' average ability levels ranged within 1/3 standard deviation (see Table A-5).

Cadet Demographics. A similar set of demographic variables was analyzed for the OTS cadets tested on the 14 Aviation Information and General Science booklets (see Tables A-1 through A-4). On the average, the cadets were 25 to 26 years of age and had completed 16 years of education. The proportions of males and females in the samples were 85% to 90% and 10% to 15%, respectively. The racial representation was 90% to 95% White, 1% to 5% Black, 1% to 2% Hispanic, and 2% or less in other categories.

Item Difficulty (Common Items). The performance of examinees on the 15 to 20 common items placed in each test booklet was evaluated by computing the mean and standard deviation of item difficulty (proportion correct) in the power subtests (see Table A-6). On the average, the proportion of airmen samples answering the common items correctly differed by .06 or less on nine subtests and by .09 or less on all 14 power subtests. The cadet samples were especially consistent, with the means across samples differing by only about .02 (.595 to .617 on Aviation Information and .520 to .541 on General Science).

Item and Test Characteristics

Further analyses were conducted by sample to evaluate the adequacy of the items in the experimental pool. Item and test indices were derived using classical or "true score" analytic techniques (Gulliksen, 1950; Koplyay, 1981).

Item Discrimination. The biserial correlation (r_{bis}) between item score (correct or incorrect) and total test score (subtest raw score) was obtained as an index of the discriminative value of each item. The analyses made possible the identification of new items that reached or exceeded standards for distinguishing among examinees of differing ability levels.

A primary requirement was that item-total score biserial correlations be negative for all non-keyed (incorrect) alternatives. Items failing to meet this standard were dropped from the pool.

A further objective for the initial pool of experimental items[2] was to have at least 100 items in each subtest with biserials of .50 or greater for the keyed (correct) alternative. This goal was met by 9 of the 16 subtests (Table 2). Five subtests substantially exceeded the goal: Hidden Figures (273), Electrical Maze (246), Instrument Comprehension (204), Block Counting (156), and Table Reading (141). Many Aviation Information (AI) and General Science (GS) items were too difficult to achieve appropriate statistics; however, after readministration to cadets, item data were significantly improved for AI (141 with biserials of .50 or greater), but indicated that more GS items needed to be created. Results for the Scale Reading (SR) subtest, a speeded test, indicated that difficulty interacted with discrimination to reduce the size of biserials. As shown in Table 2, a substantial pool of items meeting a minimal discriminative standard of $r_{bis} \geq .40$ was available for Form P test construction.

Table 2. Item Discrimination (r_{bis}) and Test Reliability (KR-20)
Data for the Experimental Item Pool

	Item discrimination		Test reliability
Subtest	No. of items $r_{bis} \geq .50$	No. of items $r_{bis} \geq .40$	Range
Verbal Analogies	44	101	.79 – .89
Arithmetic Reasoning	117	179	.86 – .92
Reading Comprehension	112	169	.87 – .90
Data Interpretation	125	186	.85 – .88
Word Knowledge	76	130	.81 – .88
Math Knowledge	97	146	.85 – .91
Mechanical Comprehension	29	85	.69 – .79
Electrical Maze	246	282	.94 – .95
Scale Reading	60	214[a]	.84 – .91[b]
Instrument Comprehension	204	250	.92 – .94
Block Counting	156	186	.90 – .95
Table Reading	141	166[a]	.95 – .96[b]
Aviation Information			
Airmen	22	69	.67 – .72
Cadets	141	176	.91 – .93
Rotated Blocks	114	178	.87 – .90
General Science			
Airmen	17	57	.61 – .78
Cadets	54	86	.76 – .86
Hidden Figures	273	284	.95 – .96

[a] The number of acceptable items include those obtaining $r \geq .30$.

[b] The items on these tests are easy, are homogeneous in type, and were administered under speeded conditions; hence, the KR-20 values are inflated.

_____

[2] Additions to this initial pool of AFOQT items were constructed in a phase of this project subsequent to the development of Form P. That development will be described in more detail in another report.

Test Reliability. With the exception of the results for the airmen samples for GS and AI, internal consistency reliability coefficients (KR-20) obtained for the experimental item pool data ranged from .69 to .96 (Table 2). The lowest (.69 - .79) were for the seven sets of Mechanical Comprehension items. The highest reliabilities (.90 to .96) were derived for the Electrical Maze, Instrument Comprehension, Block Counting, Table Reading, Aviation Information (Cadet sample), and Hidden Figures subtests. Reliabilities of .81 to .92 were obtained for Arithmetic Reasoning, Reading Comprehension, Data Interpretation, Word Knowledge, Math Knowledge, Scale Reading, and Rotated Blocks. Verbal Analogies and General Science (Cadet sample) had reliability coefficients ranging from .76 to .89. In summary, all the subtests except Mechanical Comprehension had reliability coefficients of .76 or higher. These results compare favorably with those reported for officer applicants in the equating/standardization sample for Form O (Rogers, Roach, & Wegner, 1986).

III. PARALLEL FORMS DEVELOPMENT

Technical Approach

Objectives

The construction of Form P was designed to meet two objectives: the development of alternate forms (P1 and P2) which were (a) equivalent and (b) parallel to Form O. The foremost consideration in selecting statistically qualified items from the new experimental pool was a high degree of resemblance to Form O subtests in terms of distributions of item difficulty and keyed responses, content/subject matter, item position, and stylistic features such as item length, illustrations, format, and appearance.

Item and Test Features

Item Difficulty. Early in the test construction project, concerns were raised as to how accurately item difficulty indices computed from responses of basic airmen and officer cadet subjects would reflect the actual difficulty of new items for officer applicants. The precision of item difficulty indices was questioned for two reasons. The first concerned the ability level of subjects and the second, an apparent speeded component underlying several subtests defined as power tests in AFOQT Form O (Rogers, Roach, & Wegner, 1986).

Basic airmen were anticipated to find the items more difficult on the average than would officer applicants, the majority of whom have completed more years of formal education. Conversely, the cadets, as a select group who had been previously screened and found to meet or exceed educational entry standards for the officer force as well as AFOQT score minimums, were expected to perform better than the larger pool of officer applicants.

Item-omitting rates for officer applicants tested on AFOQT Form O strongly implied that the majority of AFOQT subtests had a speeded component (see Table 4 in Rogers, Roach, & Wegner, 1986, and Appendix B in Skinner & Ree, 1987). Completion rates were often lower than 95%, the level traditionally accepted as defining an instrument as a power test. Furthermore, difficulty values tended to change systematically with item placement. Items located toward the end of selected subtests tended to have increasingly lower difficulty values. Collectively, the data suggested that computed difficulties were probably contaminated by the speeded quality of the test. Item difficulty was calculated as the proportion of officer applicants tested who chose the correct option and was not adjusted to account for examinees who failed to attempt an item within prescribed time limits. The net result was that items placed later in some AFOQT Form O subtests

appeared on the basis of computed statistics to be more difficult than if they had been tested under pure power conditions.

Together, the subject characteristic and test speededness issues posed a problem for designing parallel test forms, particularly with respect to raw score equivalence. Information was needed to identify new items for AFOQT Form P which would be comparable in difficulty to those residing in the same position of AFOQT Form O. The approach used was to estimate statistically the difficulty of each newly developed item for officer applicants. The approach taken is described later in the Analysis section.

Item Discrimination. As indicated in Part II above, one of the standards for an acceptable experimental item was that biserial r's for the keyed response be equal to or greater than .40 and that they be negative for the incorrect response alternatives. The same criteria held for the selection of items in Form P. Although a higher minimum biserial ($\geq$ .50) would have been desirable, other requirements such as distributions of difficulties and content categories restricted the possible candidate items.

Content Category/Subject Matter. If parallelism to Form O were to be maximized, the content of the Form O subtests had to be categorized and replicated. This need became evident during the experimental General Science (GS) subtest development. To plan for the same distribution of content in Form P, the GS items in Form O were classified as to type of science (e.g., astronomy, biology, chemistry), and the number of items in each classification was counted. The experimental GS items were then constructed to have the same proportional representation of science types as Form O. This matching procedure was also employed for Verbal Analogies, Arithmetic Reasoning, Reading Comprehension, Word Knowledge, Math Knowledge, Mechanical Comprehension, Scale Reading, Instrument Comprehension, and Aviation Information. In some of these -- Verbal Analogies, for example -- the classification was structural (i.e., type of item format) rather than thematic. The Hidden Figures subtest had five intrinsic graphical categories with an equal number of items in Forms O and P. The five remaining subtests did not lend themselves to categorization in sufficiently meaningful ways to justify matching. The latter subtests -- Block Counting, for example -- are innately similar in task and appearance. The item-classification breakdown is shown for applicable subtests in the last section of each table in Appendix C.

Stylistic Features. Experimental item development was exacting with respect to emulating Form O stylistic features. These included format, appearance on the page, type size and type face, illustration and legend characteristics, consistency of spelling (for example, "judgment" versus "judgement"), and consistency of mathematical notation (for example, "percent" versus "%"). Certain elements had to be improved for the sake of visual clarity. For example, it was necessary to make corrections to the original negatives of airplane illustrations in Instrument Comprehension. The airplane type was essentially the same, however.

The same care was taken to match Form O styles, feature for feature, in selecting experimental items for Form P. However, a stylistic issue emerged with respect to the difficulty of items in Data Interpretation (DI) and Mechanical Comprehension (MC). DI and MC had sets of items each associated with one illustration. An initial developmental constraint was that an illustration or table could be used only once in the seven booklets of a content area. This limitation indirectly restricted the possibility for a spread of item difficulty. If an illustration served three items and only two items met the criterion of negative biserials for non-keyed options, the whole set had to be discarded. Alternatively, a set might meet correlational criteria, but its distribution of difficulty would not quite match that of a Form O item set. Since matching Form O difficulty was a requirement for every content area, this situation led to discarding otherwise acceptable item sets. The solution to these problems was to develop new DI and MC sets and booklets such that an illustration could appear in two experimental booklets (i.e., for two different item

8

sets), thereby increasing the item pool for each illustration which in turn increases the flexibility for selecting items with appropriate psychometric features for future AFOQT forms.

Keyed Response Distribution. One of the constraints in the experimental item pool development was to obtain an approximately even distribution of A to E response options. Since the Form O common items were judged to be acceptable on all criteria except this balance, selection among new items for placement in Form P included the consideration of balancing the keyed options.

The Block Counting subtest presented a special problem. Item response options consisting of numbers had to be placed in numeric sequential order. Furthermore, there had to be an equal number of A to E options. There were five items associated with one illustration, and all items had to have negative biserials for non-keyed responses. This criterion was difficult to achieve when the block count was either the smallest possible number (A response) or the largest possible number (E response). Since the answer was transparently not the opposite, E or A respectively, those response options would not be selected and the corresponding biserial would be .00, rather than negative.

Item Selection

Common Items. The selection of Form O common items for Form P was a two-stage procedure. To provide a maximum[3] of 20 common items for each booklet of a subtest, AFHRL staff selected from Form O those items with the best item discrimination statistics. Later, a subset of these items was selected for inclusion in Form P. To construct about 50% of the content of Form P subtests (typically 12 or 13 out of 25 items) with Form O common items, AFHRL used the following selection criteria in the second stage: acceptable difficulty and discrimination statistics, balanced content (where appropriate), balanced keyed response options, and matching the mean difficulty of the common items in each Form P subtest to that of each Form O subtest. In addition, the positions of the common items in each experimental subtest were considered in the selection process in order to balance items taken from the beginning, middle, and end of each subtest.

Experimental Items. The considerations for selecting common items also governed the selection of experimental items for Form P. The characteristics of the Form O items to be replaced were particularly considered and item-by-item matches were attempted. A number of trade-offs were necessary to balance content, keys, appearance, and position, but in general, priority was given to matching item difficulty and discrimination. Plots of difficulty versus discrimination were prepared separately by subtest and by content area to facilitate item selections. Both new-item and Form O common-item data appeared on the same plot. The new items whose locations on the plots were close to those of the Form O items to be replaced became candidates for selection for Form P if their item-total score biserial correlations for keyed options satisfied a minimum criterion (usually $\geq$ .50, but in some cases, $\geq$ .40).

The issues of item difficulty and item position came into conflict at times. After a Form P subtest would be half-filled with Form O common items, the positional gaps in Form P corresponding to those items not selected from Form O were to be filled by experimental items with the same difficulty as the replaced items. Exceptions to this strategy were necessary for some of the more technical subtests. Comparison of experimental item difficulty data for airmen with difficulty data for the Form O operational sample revealed that the tests were more difficult for airmen, whose responses appeared to be at the guessing level for many items in several subtests. Some replacement item candidates that were easier than the replaced Form O items in the same test

---

[3] There were 20 common items in all experimental subtests except Mechanical Comprehension (19), Rotated Blocks (15), and Hidden Figures (15).

positions were therefore selected. For the most part, however, there were a sufficient number of positional matches to permit matching on the other variables.

Form P Test Booklet Construction. The Form P test booklets were made to be virtually the same in appearance as the Form O booklet. Bodoni typeface was used, the Form O subtest instructions were duplicated, the subtest order was preserved, and the common items were in the same position they held in the Form O subtests. The manual for administration was the same as the Form O manual except for some minor improvements.

Analysis

Item Difficulty Estimates

Multiple regression analyses were conducted to obtain weights needed to estimate how difficult new items tried out on basic airmen and cadets would be upon subsequent administration of AFOQT Form P to officer applicants. Primary data were difficulty indices for common items appearing in AFOQT Form O and in each experimental test booklet. Item difficulty indices used as elements of the criterion vector were obtained from analyses of responses of 75,980 officer applicants administered AFOQT Form O under operational testing conditions between 1 March 1982 and 29 February 1984. Two types of information were included in the predictor set. Elements of the primary predictor vector for the corresponding common items were item difficulty values computed on airmen (or cadet) responses to experimental test booklets. A second predictor variable was developed to account for the potential relationship between the difficulty of an item and its location within the subtest. The location or position of each common item was recorded as its subtest item number in AFOQT Form O.

Analyses were conducted separately for each of 14 subtests in AFOQT Form O that had been treated specifically as power tests during the development and standardization of AFOQT Form O (Rogers, Roach, & Wegner, 1986). In the General Science and Aviation Information subtests, regression analyses were repeated on data collected from officer cadets to supplement the difficulty estimates obtained from basic airmen samples. Two equations were solved for each subtest and sample combination. One model constrained the relationship between item difficulty for officer applicants and basic airmen (or cadets) by item location to a linear form. Specifications for the second model permitted the relationship to take the more complex form of a curvilinear function. The total number of elements (N) for each model was equal to the number of common items in each subtest times seven, the number of independent airmen (and cadet) samples for which common item difficulty values were available.

Inspection of squared multiple correlation coefficients (R^2) and standard errors of estimate (SEE) for the two models revealed that the data were adequately described by the more simple linear function. Results of the analyses for each subtest are shown in Appendix B. Follow-on analyses were conducted using the derived weights to compute estimates of the difficulty of new items for selected subtest positions in AFOQT Form P. These estimates were then used to aid in selection of experimental items for Form P.

Test Analyses

The analytic techniques used to identify items that would result in parallel forms included computation of mean difficulty and mean biserial r's, frequency and percentage distributions of these variables, and frequency distributions of items in content categories. Appendix C consists of a series of tables presenting these variables for each subtest (see Tables C-1 through C-16).

10

Each table displays ranges of difficulty and biserial r's at the left side. The columns refer to two samples: the 75,980 officer applicants who took the operational Form 0 and the approximately 350 airmen who took the experimental tests, including common items from Form 0 and newly developed items. (The table for Aviation Information refers to a sample of 200 cadets, rather than airmen.) The column headed "Retained" refers to those items selected from the Form 0 officer sample data to be common items in the experimental pool.

The columns on the right half of the table refer to the number and percentage of new subtest items in each difficulty and biserial range that were placed in Forms P1 and P2, and the number of common and new items combined ("Total") in each range. For example, the data for Verbal Analogies (Table C-1) show that two items in Form 0 fell in the difficulty range of .90-.99, one item of which was retained for placement as a common item in Forms P1 and P2. One new item in Form P1 and a different new item in Form P2 were in the .90-.99 difficulty range. When this new item is added to the one retained common item, it results in two items in each Form P that were in the .90-.99 difficulty range.

Mean difficulty and mean biserials are shown in rows at the bottom of their respective sections for each configuration of test items. These means and the distributions above them permit comparisons among Form 0, Form P1, and Form P2 with respect to central tendencies and distributions of difficulty and discrimination.

For applicable subtests, the last section of the table deals with content, structural, or stylistic categories. Frequencies of items in each category are shown for Form 0, the items retained from Form 0 as common items for Form P, the new items, and the sum of common and new items ("Total"). Thus, for Verbal Analogies it can be seen that each Form P distribution in terms of format was the same as the Form 0 distribution.

Results and Discussion

The emphasis of the Form P test construction effort was on achieving a high degree of similarity to Form 0. The data presented in Appendix C show that matches were achieved to a large extent for item difficulty, item discrimination, and distribution of content and stylistic features. Average item difficulty and discriminative values for power subtests (Table Reading and Scale Reading are excluded) in the three test forms have been extracted from Appendix C and are further summarized in Table 3.

Item Difficulty

Comparison of the mean difficulty of Form 0 items obtained from officer data with the data[4] for the experimental items selected for Form P shows very close matching (Appendix C), with the difficulty indices rarely differing by more than .02.[5] The mean difficulties of Arithmetic Reasoning, Reading Comprehension, Electrical Maze, and Hidden Figures were the same in Forms 0, P1, and P2. There was a .01 variation in difficulty among the forms for Verbal Analogies, Word Knowledge, Math Knowledge, Instrument Comprehension, Aviation Information, and Rotated Blocks. A variation of .02 was found among the forms for Mechanical Comprehension, Block Counting and General Science. Mean difficulties among forms for Data Interpretation and Scale Reading differed by .05 and .04, respectively, with Form P showing the higher mean difficulties.

---

[4] Airmen data were transformed to estimate officer difficulty.

[5] Appendix C does not show difficulty data for Table Reading, a speeded subtest that does not contain inherently difficult items.

11

For Forms P1 and P2 only, there were no differences in mean difficulty for 10 power subtests and differences of .01 for four power subtests (Table 3). The easiest subtests appeared to be Reading Comprehension (mean difficulty .64) and Hidden Figures (mean difficulty .63). The hardest subtests were Electrical Maze (.38), Aviation Information (.43), General Science (.44), and Instrument Comprehension (.44 to .45).

Table 3. Mean Item Difficulty and Discrimination for Power Subtests
in AFOQT Form 0 and Experimental Forms P1 and P2

	Item difficulty			Item discrimination		
Subtests	Form 0	Form P1	Form P2	Form 0	Form P1	Form P2
Verbal Analogies	.60	.61	.60	.60	.47	.47
Arithmetic Reasoning	.53	.53	.53	.60	.56	.56
Reading Comprehension	.63	.64	.64	.67	.54	.55
Data Interpretation	.48	.52	.53	.49	.45	.48
Word Knowledge	.55	.56	.56	.67	.55	.57
Math Knowledge	.57	.58	.58	.64	.55	.54
Mechanical Comprehension	.51	.50	.49	.51	.39	.39
Electrical Maze	.38	.38	.38	.63	.68	.67
Instrument Comprehension	.44	.45	.44	.63	.59	.62
Block Counting	.53	.51	.51	.65	.68	.65
Aviation Information	.44	.43	.43	.56	.59	.58
Rotated Blocks	.50	.51	.51	.46	.54	.54
General Science	.42	.44	.44	.50	.43	.40
Hidden Figures	.63	.63	.63	.61	.67	.68

Note. Form 0 values are based on an officer applicant sample. Forms P1 and P2 values are based on experimental item administrations to airmen or cadet samples.

Item Discrimination[6]

The mean keyed-response biserials ranged from .46 to .67 for the subtests in Form 0 and from .39 to .68 for those in Form P (Appendix C). The discrimination of common and new items in the Form P subtests is shown by a range of mean biserials from .34 to .68 (common items) and .40 to .73 (new items). The highest mean biserials in the total Forms P1 and P2, respectively, are seen for Hidden Figures (.67 and .68), Electrical Maze (.68 and .67), and Block Counting (.68 and .65). The lowest mean biserials in P1 and P2 are for Scale Reading (.42 and .38), General Science (.43 and .40), and Mechanical Comprehension (.39 for both forms).

Comparison of Form 0 and Form P indicates that 10 subtests yielded biserials that differed by .10 or less among the forms, and six that differed from .11 to .15. The largest differences in mean biserials between the two forms are for Table Reading (Form 0, .56; Forms P (speeded), .71 and .70), Verbal Analogies (Form 0, .60; Forms P, .47 and .47), and Reading Comprehension (Form 0, .67; Forms P, .54 and .55).

The parallelism of Forms P1 and P2 with respect to mean biserials is closer than that between Forms 0 and P. Forms P1 and P2 have the same mean biserials for five subtests, differences of .01 for five subtests, and differences from .02 to .04 for six subtests.

Content Categories

Close matches were achieved for subtest content and style categories across forms (Appendix C). Comparisons show a high frequency of zero or one-item differences in numbers per category among Forms 0, P1, and P2. As a result of decisions to improve distributions of content, only two subtests showed differences greater than two items within a single category. These were Instrument Comprehension, in which one of the 10 categories had a difference of three items between Forms 0 and P, and another had a difference of four items; and Math Knowledge, where one of its five categories had a difference of four items. In the latter case, the affected category was "arithmetic reasoning," a category which appeared in Form 0 but was dropped in Form P because the content was redundant to that of the Arithmetic Reasoning subtest.

IV. CONCLUSIONS AND RECOMMENDATIONS

The criteria and methodology for selecting common items from Form 0 and new items from the experimental pool for Form P resulted in distributions of difficulty, discrimination, and content which suggest that Forms P1 and P2 are virtually equivalent, and that Form P is highly comparable to Form 0, if not equivalent. The data strongly imply that the new forms adequately represent the psychological and psychometric properties of Form 0.

Prior to operational use, it is recommended that Form P undergo additional evaluation. Preliminary try-out administrations with military samples are needed to obtain evidence that the psychometric properties associated with the experimental administration of the selected common and new items are the same when these items appear in the new contexts of Form P. The adequacy of Form P items should be verified. Items that fail to meet the standards of difficulty and discrimination should be revised or replaced and then tried out and reanalyzed in an iterative procedure.

---

[6] The discussion of item discrimination is based on the data in Appendix C. References to Table Reading and Scale Reading are to the speeded analyses only.

Further, in order to provide continuity of score interpretation and of standards for Form O qualifying scores, development of tables for equating Form P to Form O is needed. Finally, it is recommended that Form P be field tested on a sample of officer applicants corresponding in education and ability level to the target population for the AFOQT.

REFERENCES

Air Force Regulation 35-8 (1983). Air Force military personnel testing system. Washington, DC: Department of the Air Force.

Arth, T.O. (1986). Validation of the AFOQT for non-rated officers (AFHRL-TP-85-50, AD-A164 134). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Arth, T.O., & Skinner, M.J. (1986). Aptitude selectors for Air Force officer non-aircrew jobs. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 301-306), New London, CT.

Arth, T.O., Steuck, K.W., Sorrentino, C.T. & Burke, E.F. (in preparation). Air Force Officer Qualifying Test (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Finegold, L.S., & Rogers, D.L. (1985). Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training (AFHRL-TR-85-13, AD-A158 162). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley and Sons, Inc.

Koplyay, J.B. (1981). Item analysis program (IAP) for achievement tests (AFHRL-TP-81-22, AD-A107 884). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogers, D.L., Roach, B.W., & Short, L.O. (1986). Mental ability testing in the selection of Air Force officers: A brief historical overview (AFHRL-TP-86-23, AD-A173 484). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogers, D.L., Roach, B.W., & Wegner, T.G. (1986). Air Force Officer Qualifying Test Form O: Development and standardization (AFHRL-TR-86-24, AD-A172 037). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Skinner, M.J., & Ree, M.J. (1987). Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O (AFHRL-TR-86-68, AD-A184 975). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

United States Air Force (1987). Air Force Officer Qualifying Test information pamphlet (AFPT 997). Randolph AFB, TX: Air Force Military Personnel Center.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

APPENDIX A. SAMPLE COMPARABILITY ANALYSES: SUMMARY STATISTICS FOR
DEMOGRAPHIC AND PERFORMANCE VARIABLES

Table A-1. Age of Examinees by Sample

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	M	19.9	20.1	19.6	20.1	20.1	20.0	20.0
	SD	2.3	2.3	2.0	1.9	2.1	2.1	2.0
Arithmetic Reasoning	M	19.9	20.1	19.6	20.4	20.2	20.2	20.4
	SD	2.2	2.2	1.9	2.0	2.1	2.1	2.0
Reading Comprehension	M	19.3	19.3	19.2	19.1	18.9	19.4	19.0
	SD	2.1	2.2	2.0	1.9	1.8	2.3	1.8
Data Interpretation	M	19.0	19.3	19.1	19.5	19.5	19.3	19.3
	SD	1.9	2.3	1.9	2.1	2.1	2.1	2.0
Word Knowledge	M	18.9	19.1	19.2	19.5	19.0	19.2	18.9
	SD	1.7	2.0	2.1	2.3	1.8	2.0	1.8
Math Knowledge	M	19.2	19.5	19.5	19.6	19.6	19.7	19.8
	SD	2.0	2.1	2.2	2.1	2.0	2.0	2.2
Mechanical Comprehension	M	19.0	19.1	19.2	19.2	19.6	19.8	19.7
	SD	1.8	1.9	2.0	2.0	2.0	2.2	2.2
Electrical Maze	M	18.8	19.3	19.1	19.6	19.7	19.8	19.7
	SD	1.7	2.0	1.9	2.0	2.0	2.2	2.0
Scale Reading (Forward Order)	M	19.1	19.1	19.1	18.8	18.7	18.8	18.9
	SD	2.1	2.0	2.1	1.8	1.8	1.8	2.0
Scale Reading (Reverse Order)	M	19.3	19.0	19.1	19.4	19.5	19.3	19.4
	SD	2.0	1.8	1.9	2.0	2.2	2.0	2.2
Instrument Comprehension	M	19.6	19.9	19.9	20.0	20.1	20.0	20.0
	SD	1.9	2.1	1.9	1.9	2.1	2.1	2.0
Block Counting	M	19.1	19.6	19.5	19.6	19.8	19.9	20.1
	SD	2.0	2.2	2.2	2.1	2.0	2.2	2.3
Table Reading (Forward Order)	M	20.2	20.3	20.3	20.2	20.3	20.5	19.3
	SD	2.0	2.1	2.0	2.0	2.2	2.3	2.1
Table Reading (Reverse Order)	M	20.1	20.3	20.3	20.4	20.1	20.5	19.4
	SD	2.0	2.2	2.0	2.1	1.8	2.2	2.1
Aviation Information Airmen	M	20.1	20.1	20.0	20.0	20.3	20.2	20.2
	SD	1.9	2.1	2.1	2.0	2.0	2.1	2.2
Aviation Information Cadets	M	25.4	25.3	25.6	25.5	25.6	25.9	25.5
	SD	2.8	2.8	3.0	2.8	2.9	3.0	3.1

17

Table A-1. (Concluded)

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Rotated Blocks	M	19.1	19.1	19.3	19.2	19.6	19.8	19.7
	SD	1.9	1.9	2.0	2.1	2.0	2.2	2.2
General Science Airmen	M	20.3	20.6	20.1	20.6	18.7	18.7	18.9
	SD	2.2	2.3	1.8	2.3	1.8	1.8	1.9
General Science Cadets	M	25.4	25.5	25.6	25.7	25.4	25.6	25.6
	SD	2.9	3.0	3.0	2.7	3.0	2.9	3.3
Hidden Figures	M	18.9	19.1	19.2	19.6	19.0	19.2	18.9
	SD	1.6	2.0	2.1	2.3	1.9	1.9	1.8

Note. Airmen samples except where cadet samples are noted.

Table A-2. Education Level (in years) of Examinees by Sample

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	M	12.6	12.6	12.4	12.6	12.5	12.5	12.5
	SD	1.2	1.1	.8	1.0	1.0	1.0	1.0
Arithmetic Reasoning	M	12.5	12.6	12.4	12.7	12.7	12.7	12.6
	SD	1.2	1.1	.8	1.1	1.2	1.1	1.0
Reading Comprehension	M	12.4	12.4	12.4	12.3	12.4	12.4	12.4
	SD	.9	.9	1.0	.8	.9	1.0	.8
Data Interpretation	M	12.4	12.4	12.4	12.5	12.4	12.4	12.4
	SD	.9	1.0	.9	1.1	.9	.9	1.0
Word Knowledge	M	12.3	12.3	12.4	12.5	12.4	12.4	12.4
	SD	1.0	.8	1.0	1.0	.9	.9	1.1
Math Knowledge	M	12.4	12.4	12.5	12.4	12.4	12.4	12.4
	SD	.9	.9	1.2	1.1	.8	.9	.9
Mechanical Comprehension	M	12.3	12.3	12.3	12.4	12.5	12.4	12.4
	SD	.8	.9	.8	.9	1.1	.9	.9
Electrical Maze	M	12.2	12.4	12.3	12.4	12.4	12.5	12.5
	SD	.6	.9	.7	.8	.9	.9	1.0
Scale Reading (Forward Order)	M	12.4	12.4	12.4	12.3	12.2	12.2	12.2
	SD	.9	.8	.9	.7	.8	.6	.7
Scale Reading (Reverse Order)	M	12.4	12.3	12.3	12.5	12.4	12.4	12.4
	SD	.9	.8	.9	.9	.9	1.1	1.0
Instrument Comprehension	M	12.3	12.4	12.5	12.5	12.5	12.5	12.5
	SD	.8	1.0	1.0	1.0	1.0	1.0	1.0
Block Counting	M	12.4	12.4	12.5	12.5	12.5	12.6	12.6
	SD	.9	.9	1.1	1.1	1.0	1.3	1.1
Table Reading (Forward Order)	M	12.6	12.6	12.7	12.6	12.6	12.5	12.4
	SD	1.0	1.1	1.1	1.1	1.0	1.0	.9
Table Reading (Reverse Order)	M	12.7	12.6	12.6	12.6	12.6	12.6	12.5
	SD	1.1	1.1	1.1	1.0	1.0	1.1	.9
Aviation Information Airmen	M	12.6	12.6	12.5	12.5	12.6	12.6	12.6
	SD	1.0	1.0	1.1	1.0	1.0	1.1	1.0
Aviation Information Cadets	M	16.3	16.2	16.2	16.2	16.2	16.2	16.2
	SD	.6	.6	.7	.5	.5	.7	.6

19

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Rotated Blocks	M	12.3	12.3	12.3	12.4	12.4	12.4	12.4
	SD	.9	.8	.8	1.0	.9	.8	.9
General Science Airmen	M	12.7	12.6	12.6	12.7	12.3	12.2	12.2
	SD	1.1	1.0	1.0	1.1	.8	.9	.6
General Science Cadets	M	16.3	16.2	16.3	16.2	16.2	16.3	16.2
	SD	.8	.6	.6	.6	.6	.8	.5
Hidden Figures	M	12.3	12.3	12.4	12.5	12.3	12.4	12.4
	SD	.9	.9	1.0	.9	.9	.9	1.0

Note. Airmen samples except where cadet samples are noted.

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	M	63.6	54.7	93.0	81.1	83.1	88.6	73.3
	F	36.4	45.3	7.0	18.9	16.9	11.4	26.7
Arithmetic Reasoning	M	64.2	55.4	93.0	100.0	100.0	100.0	100.0
	F	35.8	44.6	7.0	0.0	0.0	0.0	0.0
Reading Comprehension	M	95.1	95.0	95.1	89.9	88.7	86.0	80.0
	F	4.9	5.0	4.9	10.1	11.3	14.0	20.0
Data Interpretation	M	87.2	82.9	76.7	81.9	86.9	91.9	92.3
	F	12.8	17.1	23.3	18.1	13.1	8.1	7.7
Word Knowledge	M	100.0	85.6	63.8	89.6	96.5	78.8	95.7
	F	0.0	14.4	36.2	10.4	3.6	21.2	4.3
Math Knowledge	M	71.2	100.0	77.0	97.6	92.4	82.8	94.8
	F	28.8	0.0	23.0	2.4	7.7	17.2	5.2
Mechanical Comprehension	M	100.0	100.0	99.4	92.5	93.6	81.1	75.1
	F	0.0	0.0	.6	7.5	6.4	18.9	24.9
Electrical Maze	M	86.9	82.8	85.4	92.5	83.0	95.0	82.7
	F	13.1	17.3	14.6	7.5	17.0	5.0	17.4
Scale Reading (Forward Order)	M	100.0	100.0	100.0	97.2	97.4	93.4	93.8
	F	0.0	0.0	0.0	2.9	2.6	6.7	6.2
Scale Reading (Reverse Order)	M	92.5	100.0	91.8	84.1	85.0	90.2	91.7
	F	7.5	0.0	8.2	16.0	15.0	9.8	8.3
Instrument Comprehension	M	92.9	90.6	100.0	80.7	83.2	88.4	73.3
	F	7.1	9.4	0.0	19.3	16.8	11.6	26.7
Block Counting	M	71.1	100.0	76.4	97.6	82.5	63.5	54.5
	F	29.0	0.0	23.6	2.4	17.5	36.5	45.5
Table Reading (Forward Order)	M	82.7	85.3	95.9	95.9	100.0	100.0	93.8
	F	17.3	14.7	4.2	4.1	0.0	0.0	6.2
Table Reading (Reverse Order)	M	88.9	86.5	96.2	96.2	100.0	100.0	93.4
	F	11.1	13.5	3.8	3.8	0.0	0.0	6.6
Aviation Information Airmen	M	81.0	83.7	88.7	74.6	78.3	82.6	83.1
	F	19.0	16.3	11.3	25.4	21.8	17.4	16.9
Aviation Information Cadets	M	87.9	90.3	87.0	84.0	89.5	84.8	86.1
	F	12.1	9.7	13.0	16.0	10.5	15.2	13.9

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Rotated Blocks	M	100.0	100.0	99.2	92.0	93.4	81.0	76.3
	F	0.0	0.0	.8	8.0	6.6	19.1	23.7
General Science Airmen	M	100.0	100.0	100.0	100.0	96.3	93.2	92.6
	F	0.0	0.0	0.0	0.0	3.8	6.8	7.4
General Science Cadets	M	86.3	88.0	87.9	86.9	89.8	85.5	85.1
	F	13.7	12.0	12.1	13.1	10.2	14.2	14.9
Hidden Figures	M	100.0	85.4	63.5	89.8	96.5	78.4	95.9
	F	0.0	14.6	36.5	10.2	3.5	21.6	4.1

Note. Airmen samples except where cadet samples are noted. Sample percentages may not sum to 100.0 due to rounding.

Ethnicity Composition of Examinees by Sample (in Percentages)

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	I	0.0	.6	.6	.6	1.5	.6	.6
	A	2.3	.9	5.4	1.2	1.5	2.4	2.3
	B	16.7	21.0	16.4	17.5	14.2	12.5	16.1
	H	4.8	4.5	7.1	6.9	2.6	6.8	4.3
	W	76.2	73.1	70.5	73.9	80.4	77.7	76.8
Arithmetic Reasoning	I	0.0	.6	.6	.3	.3	.9	.6
	A	2.3	.9	5.4	2.6	3.8	2.0	2.0
	B	17.4	20.0	16.6	14.0	14.5	12.9	14.0
	H	4.9	4.5	7.0	6.3	2.7	6.9	6.3
	W	75.4	74.1	70.4	76.8	78.8	77.3	77.1
Reading Comprehension	I	1.2	.6	.3	.6	.3	.8	.6
	A	1.2	2.1	1.4	2.9	2.0	1.7	2.0
	B	10.7	13.8	12.0	12.5	15.4	14.0	13.1
	H	6.4	2.9	6.3	5.5	4.9	6.2	4.9
	W	80.6	80.6	80.0	78.5	77.3	77.3	79.4
Data Interpretation	I	.8	.8	.8	.5	1.1	.3	.3
	A	3.6	2.2	2.2	2.3	1.6	1.3	1.6
	B	11.8	13.0	16.2	19.0	18.1	14.0	14.9
	H	6.9	3.9	5.9	4.7	3.2	7.0	5.8
	W	76.9	80.1	74.9	73.4	76.1	77.4	77.5
Word Knowledge	I	.3	.3	.0	.9	.3	.0	1.2
	A	1.5	1.5	1.2	1.8	2.9	2.4	1.5
	B	9.3	10.9	16.4	9.4	9.3	12.7	13.3
	H	2.7	3.2	4.6	4.1	6.4	2.4	3.2
	W	86.2	84.2	77.8	83.8	81.1	82.5	80.7
Math Knowledge	I	.3	.9	.6	.3	.3	.9	.9
	A	2.3	1.7	1.4	3.0	2.5	4.3	2.3
	B	13.7	12.1	14.9	8.1	14.2	16.1	15.6
	H	1.8	3.2	4.9	4.1	4.1	6.0	5.2
	W	81.9	82.1	78.2	84.6	79.0	72.7	76.1
Mechanical Comprehension	I	.6	.8	.8	.3	.0	1.1	.9
	A	1.9	3.3	1.4	1.1	3.3	3.9	1.1
	B	13.3	16.6	14.4	15.0	19.4	17.2	14.9
	H	4.7	3.3	5.3	5.3	4.2	4.7	5.4
	W	79.6	76.0	78.1	78.4	73.1	73.1	77.7
Electrical Maze	I	.8	.3	.9	.6	.9	1.2	2.1
	A	1.3	1.2	.9	2.2	3.8	2.0	3.3
	B	11.7	10.2	14.8	14.1	16.4	15.7	13.8
	H	5.1	5.6	4.1	3.9	5.8	5.3	3.6
	W	81.1	82.8	79.4	79.2	73.2	75.8	77.3

Subtest		1	2	3	4	5	6	7
					Sample/Booklet Number			
Scale Reading	I	1.4	.3	.3	.6	2.0	.3	.9
(Forward Order)	A	.9	1.5	1.1	2.0	1.7	.6	1.1
	B	9.7	11.8	12.4	13.4	10.4	12.4	13.1
	H	6.3	3.0	7.1	3.7	4.1	4.3	3.7
	W	81.8	83.5	79.1	80.4	81.8	82.4	81.1
Scale Reading	I	.6	.6	.3	.6	1.1	.3	.3
(Reverse Order)	A	2.2	2.5	1.1	2.3	1.7	.9	1.7
	B	14.4	15.5	15.4	18.1	18.7	14.7	14.1
	H	2.8	5.0	6.3	4.9	2.8	6.5	5.2
	W	80.1	76.5	76.9	74.1	75.6	77.8	78.7
Instrument Comprehension	I	.6	.3	.0	.6	1.4	.9	.6
	A	5.4	1.9	.8	1.2	1.4	2.1	2.3
	B	16.4	15.8	17.4	17.2	14.0	13.1	16.5
	H	6.8	3.1	3.3	7.2	2.6	6.9	4.4
	W	70.8	78.9	78.6	73.9	80.5	77.0	76.2
Block Counting	I	.3	.6	.9	.0	2.1	.0	.6
	A	2.0	2.0	1.4	2.7	2.9	2.3	.9
	B	13.3	11.6	15.2	8.6	14.4	17.1	20.3
	H	1.7	3.2	4.9	4.3	3.5	4.9	4.8
	W	82.6	82.6	77.6	84.5	77.1	75.7	73.5
Table Reading	I	.6	.3	.3	.3	.8	.6	.0
(Forward Order)	A	2.6	2.3	2.6	3.8	1.9	2.8	1.2
	B	12.5	13.7	15.2	14.5	13.3	11.3	14.4
	H	6.0	4.6	7.0	2.6	4.4	5.0	5.7
	W	78.4	79.2	74.9	78.8	79.6	80.4	78.6
Table Reading	I	.6	1.2	.6	1.2	1.7	.6	1.0
(Reverse Order)	A	2.0	2.6	2.3	1.2	3.6	2.5	2.3
	B	12.9	13.5	11.1	15.5	14.3	14.7	11.6
	H	5.8	6.0	6.1	5.9	4.2	6.9	4.8
	W	78.7	76.8	79.9	76.3	76.2	75.3	80.4
Aviation Information Airmen	I	.6	1.4	.9	.6	.6	.3	1.1
	A	1.2	1.4	2.4	2.3	1.7	2.8	2.3
	B	17.5	14.0	13.2	16.4	13.9	11.3	14.0
	H	7.2	2.6	6.9	4.1	5.7	6.5	5.1
	W	73.6	80.6	76.7	76.6	78.1	79.0	77.4
Aviation Information Cadets	I	1.0	.0	.5	1.5	1.0	.0	.0
	A	1.5	1.0	3.9	2.9	2.4	1.4	2.2
	B	4.4	2.4	1.9	3.4	3.9	1.0	1.1
	H	1.9	2.4	1.9	1.9	2.4	1.0	1.6
	W	91.3	94.2	91.8	90.3	90.4	96.7	95.1

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Rotated Blocks	I	.6	.6	1.1	.3	.0	1.1	1.1
	A	2.2	3.3	1.4	1.1	3.0	3.7	1.1
	B	13.3	16.5	15.9	14.3	20.2	16.3	14.7
	H	4.4	3.6	4.7	5.5	4.1	4.8	5.1
	W	79.5	76.1	76.9	78.8	72.7	74.2	77.9
General Science	I	.8	.6	1.7	.6	1.2	1.1	1.1
Airmen	A	2.0	2.8	3.6	2.8	3.2	.8	.6
	B	13.7	10.6	13.1	14.7	11.2	10.4	12.3
	H	4.5	5.0	4.5	7.0	4.6	5.1	4.3
	W	79.1	81.0	77.1	74.9	79.8	82.6	81.8
General Science	I	.0	1.0	1.0	.5	1.0	.0	.0
Cadets	A	2.5	1.0	2.6	3.1	2.0	2.4	1.0
	B	5.5	1.5	2.1	2.1	2.5	4.4	1.5
	H	2.5	1.0	1.5	1.6	1.5	2.4	2.5
	W	89.6	95.4	92.8	92.7	93.1	90.8	95.0
Hidden Figures	I	.3	.3	.0	1.5	.3	.0	1.2
	A	1.5	1.5	1.2	1.7	2.7	2.7	1.5
	B	8.8	11.9	16.7	9.3	9.2	11.8	14.5
	H	2.7	3.2	4.6	3.8	6.2	2.4	3.2
	W	86.7	83.2	77.6	83.8	81.7	83.1	79.7

Note. Airmen samples except where cadet samples are noted. Sample percentages may not sum to 100.0 due to rounding. I = American Indian or Alaskan Native; A = Asian or Pacific Islander; B = Black; H = Hispanic; W = White.

Table A-5. Armed Forces Qualification Test (AFQT)
Basic Airmen Percentile Scores by Sample

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	M	65.0	63.2	59.1	62.1	62.0	64.3	60.7
	SD	18.3	17.7	18.3	18.8	17.9	18.4	16.9
Arithmetic Reasoning	M	64.9	63.4	59.2	63.3	61.6	65.4	62.4
	SD	18.4	17.6	18.2	18.7	18.5	18.6	18.5
Reading Comprehension	M	62.3	59.7	60.1	61.5	60.1	61.2	59.9
	SD	17.4	17.4	18.1	17.3	18.3	18.0	17.9
Data Interpretation	M	61.5	60.5	60.0	61.5	61.1	60.9	62.4
	SD	17.4	18.0	17.8	18.2	17.5	17.7	17.8
Word Knowledge	M	65.0	62.7	63.7	64.1	64.2	67.3	63.2
	SD	16.4	15.7	16.7	18.0	16.0	16.9	16.9
Math Knowledge	M	64.6	64.8	64.8	63.0	64.7	62.1	63.4
	SD	16.4	17.6	16.3	17.5	17.0	18.4	18.5
Mechanical Comprehension	M	60.0	60.2	59.9	61.0	59.8	60.2	63.0
	SD	18.0	17.4	17.2	18.0	18.1	17.8	17.2
Electrical Maze	M	64.0	65.7	63.5	64.8	62.2	63.2	63.4
	SD	15.7	17.2	16.0	17.0	18.5	18.7	19.2
Scale Reading (Forward Order)	M	60.6	60.5	60.9	59.1	58.6	60.7	61.8
	SD	17.9	18.1	18.5	17.2	17.6	17.7	18.4
Scale Reading (Reverse Order)	M	59.5	60.2	60.8	61.5	61.4	61.0	62.7
	SD	17.6	17.1	18.3	18.3	17.5	17.5	17.8
Instrument Comprehension	M	59.6	61.9	62.5	62.2	62.0	63.8	60.9
	SD	18.3	17.5	17.8	18.8	17.8	18.3	17.0
Block Counting	M	64.2	64.9	64.8	62.7	64.1	64.9	63.4
	SD	16.6	17.6	16.4	17.3	19.0	18.4	17.7
Table Reading (Forward Order)	M	62.4	63.7	62.0	62.0	62.7	63.6	61.1
	SD	18.6	18.8	18.4	18.1	18.5	18.0	17.6
Table Reading (Reverse Order)	M	63.8	62.2	64.5	61.9	62.9	63.8	61.3
	SD	18.3	18.1	18.6	18.7	18.9	19.1	17.5
Aviation Information	M	62.1	61.6	64.4	60.9	62.2	63.9	62.3
	SD	18.8	17.8	18.2	17.0	18.4	18.4	18.1
Rotated Blocks	M	60.1	60.4	60.2	61.3	60.2	60.3	63.1
	SD	17.8	17.6	18.0	17.8	18.1	17.9	17.4

Table A-5. (Concluded)

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
General Science	M	62.9	63.8	62.6	63.9	59.3	59.6	60.6
	SD	18.6	18.1	18.9	19.1	17.5	17.0	18.9
Hidden Figures	M	65.0	62.6	63.8	64.4	64.1	66.7	63.3
	SD	16.5	15.9	16.8	17.9	16.3	17.1	16.8

Table A-6. Mean Common Item Difficulty Indices for Power Subtests by Sample

Subtest		Sample/Booklet Number						
		1	2	3	4	5	6	7
Verbal Analogies	M	0.528	0.498	0.455	0.489	0.459	0.483	0.445
	SD	0.213	0.205	0.216	0.212	0.213	0.206	0.215
Arithmetic Reasoning	M	0.382	0.363	0.369	0.381	0.374	0.409	0.376
	SD	0.145	0.146	0.149	0.150	0.149	0.162	0.143
Reading Comprehension	M	0.421	0.382	0.382	0.406	0.403	0.428	0.411
	SD	0.084	0.100	0.090	0.079	0.095	0.091	0.085
Data Interpretation	M	0.431	0.432	0.443	0.463	0.448	0.434	0.450
	SD	0.192	0.186	0.183	0.168	0.193	0.196	0.193
Word Knowledge	M	0.405	0.376	0.379	0.396	0.385	0.388	0.396
	SD	0.179	0.168	0.178	0.186	0.181	0.184	0.187
Math Knowledge	M	0.356	0.386	0.352	0.363	0.320	0.350	0.349
	SD	0.122	0.133	0.127	0.138	0.131	0.138	0.126
Mechanical Comprehension	M	0.335	0.357	0.334	0.348	0.359	0.338	0.332
	SD	0.121	0.122	0.117	0.107	0.128	0.104	0.115
Electrical Maze	M	0.573	0.620	0.595	0.535	0.596	0.584	0.595
	SD	0.112	0.103	0.101	0.111	0.109	0.105	0.114
Instrument Comprehension	M	0.516	0.549	0.567	0.545	0.529	0.540	0.481
	SD	0.120	0.126	0.112	0.112	0.109	0.104	0.110
Block Counting	M	0.825	0.843	0.819	0.840	0.818	0.802	0.780
	SD	0.142	0.125	0.114	0.135	0.129	0.144	0.168
Aviation Information Airmen	M	0.297	0.322	0.302	0.302	0.305	0.316	0.330
	SD	0.123	0.127	0.106	0.115	0.114	0.113	0.124
Aviation Information Cadets	M	0.595	0.607	0.600	0.612	0.599	0.617	0.611
	SD	0.136	0.138	0.141	0.150	0.141	0.132	0.140
Rotated Blocks	M	0.421	0.456	0.404	0.439	0.452	0.426	0.423
	SD	0.193	0.185	0.178	0.193	0.196	0.194	0.174
General Science Airmen	M	0.311	0.307	0.321	0.307	0.327	0.318	0.314
	SD	0.151	0.145	0.159	0.144	0.149	0.152	0.148
General Science Cadets	M	0.535	0.535	0.533	0.522	0.540	0.520	0.541
	SD	0.174	0.184	0.185	0.180	0.174	0.176	0.178
Hidden Figures	M	0.676	0.673	0.719	0.733	0.718	0.708	0.715
	SD	0.158	0.154	0.127	0.120	0.129	0.147	0.127

Note. Airmen samples except where cadet samples are noted.

APPENDIX B.   OFFICER ITEM DIFFICULTY ESTIMATES

Table B-1. Regression Analysis Results for Predicting Item Difficulty for Officer Applicants

Subtest	Raw regression weights			R^2	SEE
	Constant	Item difficulty	Item position		
Verbal Analogies	.336976	.767917	-.007579	.83	.09
Arithmetic Reasoning	.257121	.967170	-.007078	.87	.07
Reading Comprehension	.408797	.785401	-.008505	.72	.06
Data Interpretation	.387264	.696786	-.015164	.81	.09
Word Knowledge	.385957	.642522	-.006076	.81	.07
Math Knowledge	.431834	.640306	-.004592	.72	.06
Mechanical Comprehension	.079527	1.010351	.008307	.69	.08
Electrical Maze	.276946	.589683	-.022614	.87	.07
Instrument Comprehension	.322572	.596583	-.019919	.72	.06
Block Counting	.213189	.819473	-.034267	.87	.08
Aviation Information	.151989	.999527	-.002567	.80	.06
	(-.087193)	(.884420)	(-.001303)	(.85)	(.05)
Rotated Blocks	.033552	1.021232	.003256	.96	.04
General Science	.112028	.934539	.001727	.85	.06
	(.050163)	(.750858)	(-.002430)	(.88)	(.05)
Hidden Figures	-.074246	1.165829	-.014815	.87	.08

Note. Values reported in parentheses for Aviation Information and General Science subtests are based on cadet samples. Other values are for basic airmen samples.

APPENDIX C: ITEMS SELECTED FOR AFOQT FORMS P1 AND P2 FROM FORM O AND FROM NEW ITEM POOL

Notes

1. In Tables C-1 through C-16, the Airmen columns in the "Biserial Range" section refer to the number of items associated with the given biserial correlation ranges from the experimental test data for Airmen. The "common" items are Form O items included in each of seven experimental test booklets with the same test title; the "retained" items are those common items selected for Forms P1 and P2.

2. The item-total score correlations computed for the Form O common items are represented by the mean value obtained from seven sets of items in each content area, with exceptions noted in table footnotes.

Table C-1. Difficulty Range, Biserial Range, and Format Style for Verbal Analogies

Difficulty range	Officer[a] data Form 0 Total N	Officer[a] data Form 0 Total %	Officer[a] data Form 0 Retained N	P1 New N	P2 New N	P1 Total N	P1 Total %	P2 Total N	P2 Total %
.90 - .99	2	9	1	1	1	2	8	2	8
.80 - .89	4	18	3	1	1	4	16	4	16
.70 - .79	3	14	1	2	2	3	12	3	12
.60 - .69	3	14	1	2	2	3	12	3	12
.50 - .59	1	5	1	2	2	3	12	3	12
.40 - .49	4	18	4	3	3	7	28	7	28
.30 - .39	4	18	1	1	1	2	8	2	8
.20 - .29	1	5	1	0	0	1	4	1	4
Mean difficulty	.60		.60	.62	.61	.61		.60	

Biserial range	Officer total	Airmen Common	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.70 - .79	4	0	0	0	0	0	0
.60 - .69	5	2	2	0	3	2	5
.50 - .59	10	5	2	4	2	6	4
.40 - .49	3	7	6	8	7	14	13
.30 - .39	0	6	3	0	0	3	3
Mean biserial	.60	.45	.45	.49	.50	.47	.47

Format/Style	Number in: Form 0	Retained	New	Total
One-word response choices	19	11	8	19
Two-word relationship response choices	6	2	4	6

[a] Officer Form 0 item statistics are based on test results after deletion of three VA items.

Table C-2. Difficulty Range, Biserial Range, and Content Categories for Arithmetic Reasoning

Difficulty range	Officer[a] data Form O			Officer difficulty estimated from Airmen data					
	Total		Retained	P1 New	P2 New	P1 Total		P2 Total	
	N	%	N	N	N	N	%	N	%
.80 – .89	2	10	1	1	1	2	8	2	8
.70 – .79	3	14	2	2	2	4	16	4	16
.60 – .69	3	14	2	2	2	4	16	4	16
.50 – .59	4	19	2	2	2	4	16	4	16
.40 – .49	3	14	2	2	2	4	16	4	16
.30 – .39	1	5	1	2	2	3	12	3	12
.20 – .29	5	24	2	2	2	4	16	4	16
Mean difficulty	.53		.54	.52	.52	.53		.53	

Biserial range	Officer[a] total	Airmen		P1 New	P2 New	P1 Total	P2 Total
		Common	Retained				
.70 – .79	1	1	1	2	2	3	3
.60 – .69	10	1	0	5	2	5	2
.50 – .59	7	9	6	5	8	11	14
.40 – .49	3	5	2	1	1	3	3
.30 – .39	0	2	1	0	0	1	1
.20 – .29	0	2	2	0	0	2	2
Mean biserial	.60	.48	.54	.59	.58	.56	.56

Content categories	Number in: Form O	Retained	New	Total
Geometry	5	3	3	6
Percentages	5	2	3	5
Rates / distances	10	5	4	9
Composition	2	1	1	2
Relationships	2	1	2	3

[a] Officer Form O item statistics are based on test results after deletion of four AR items.

Table C-3. Difficulty Range, Biserial Range, and Content Categories for Reading Comprehension

Difficulty range	Officer data Form 0			Officer difficulty estimated from Airmen data					
	Total		Retained	P1 New	P2 New	P1 Total		P2 Total	
	N	%	N	N	N	N	%	N	%
.80 – .89	2	8	1	1	1	2	8	2	8
.70 – .79	8	32	4	3	3	7	28	7	28
.60 – .69	4	16	2	3	3	5	20	5	20
.50 – .59	6	24	4	2	2	6	24	6	24
.40 – .49	4	16	2	3	3	5	20	5	20
.30 – .39	1	4	0	0	0	0	0	0	0
Mean difficulty	.63		.64	.63	.64	.64		.64	

Biserial range	Officer total	Airmen Common[a]	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.80 – .89	3	0	0	0	0	0	0
.70 – .79	6	0	0	0	0	0	0
.60 – .69	9	4	4	1	1	5	5
.50 – .59	6	7	5	5	6	10	11
.40 – .49	1	5	3	5	5	8	8
.30 – .39	0	4	1	1	0	2	1
Mean biserial	.67	.51	.57	.50	.53	.54	.55

Content categories	Number in: Form 0	Retained	New	Total
Life Science	1	0	2	2
Physical Science	6	4	3	7
Social Science	12	8	2	10
Art & Literature	6	1	5	6

[a] The mean biserial shown for Form 0 common items was computed from experimental Sample 1 results only.

Table C-4. Difficulty Range and Biserial Range for Data Interpretation

Difficulty range	Officer[a] data Form 0 Total N	Total %	Retained N	Officer difficulty estimated from Airmen data P1 New N	P2 New N	P1 Total N	P1 Total %	P2 Total N	P2 Total %
.80 - .89	2	9	2	0	0	2	8	2	8
.70 - .79	1	4	1	1	1	2	8	2	8
.60 - .69	3	13	2	2	2	4	16	4	16
.50 - .59	4	17	1	3	3	4	16	4	16
.40 - .49	4	17	2	3	4	5	20	6	24
.30 - .39	6	26	4	3	2	7	28	6	24
.20 - .29	2	9	0	1	1	1	4	1	4
.10 - .19	1	4	0	0	0	0	0	0	0
Mean difficulty	.48		.55	.49	.50	.52		.53	

Biserial range	Officer[a] total	Airmen Common[b]	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.70 - .79	0	0	0	0	3	0	3
.60 - .69	2	0	0	2	2	2	2
.50 - .59	10	3	1	5	3	6	4
.40 - .49	6	9	7	4	4	11	11
.30 - .39	5	1	1	2	1	3	2
.20 - .29	0	3	3	0	0	3	3
.10 - .19	0	4	0	0	0	0	0
Mean biserial	.49	.36	.40	.50	.56	.45	.48

There were no content categories for Data Interpretation.

[a] Officer Form 0 item statistics are based on test results after deletion of two DI items.

[b] The mean biserial shown for Form 0 common items was computed from experimental Sample 1 results only.

Table C-5. Difficulty Range, Biserial Range, and Parts of Speech for Word Knowledge

Difficulty range	Officer[a] data Form 0			Officer difficulty estimated from Airmen data					
	Total		Retained	P1 New	P2	P1 Total		P2	
	N	%	N	N	N	N	%	N	%
.80 - .89	1	4	1	0	0	1	4	1	4
.70 - .79	6	25	3	3	3	6	24	6	24
.60 - .69	2	8	1	1	1	2	8	2	8
.50 - .59	6	25	3	4	4	7	28	7	28
.40 - .49	4	17	2	4	4	6	24	6	24
.30 - .39	4	17	3	0	0	3	12	3	12
.20 - .29	1	4	0	0	0	0	0	0	0
Mean difficulty	.55		.55	.56	.56	.56		.56	

Biserial range	Officer[a] total	Airmen		P1 New	P2	P1 Total	P2
		Common	Retained				
.80 - .89	1	0	0	0	0	0	0
.70 - .79	8	0	0	0	3	0	3
.60 - .69	12	4	3	3	3	6	6
.50 - .59	1	10	7	8	5	15	12
.40 - .49	2	4	3	1	1	4	4
.30 - .39	0	1	0	0	0	0	0
.20 - .29	0	0	0	0	0	0	0
.10 - .19	0	1	0	0	0	0	0
Mean biserial	.67	.51	.53	.56	.61	.55	.57

Parts of Speech	Number in: Form 0	Retained	P1 New	P2	P1 Total	P2
Adjective	12	8	5	5	13	13
Noun	7	2	5	4	7	6
Verb	6	3	2	3	5	6

[a] Officer Form 0 item statistics are based on test results after one WK item was deleted.

36

Table C-6. Difficulty Range, Biserial Range, and Content Categories for Math Knowledge

| Difficulty range | Officer data Form 0 | | | Officer difficulty estimated from Airmen data | | | | | |
| | Total | | Retained | P1 New | P2 New | P1 Total | | P2 Total | |
	N	%	N	N	N	N	%	N	%
.80 – .89	1	4	0	1	1	1	4	1	4
.70 – .79	5	20	3	2	2	5	20	5	20
.60 – .69	3	12	1	2	2	3	12	3	12
.50 – .59	7	28	4	4	4	8	32	8	32
.40 – .49	8	32	4	4	4	8	32	8	32
.30 – .39	1	4	0	0	0	0	0	0	0
Mean difficulty	.57		.57	.58	.58	.58		.58	

Biserial range	Officer Total	Airmen Common	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.80 – .89	1	0	0	0	0	0	0
.70 – .79	8	1	1	2	1	3	2
.60 – .69	8	3	3	3	1	6	4
.50 – .59	5	6	2	7	10	9	12
.40 – .49	3	9	5	0	0	5	5
.30 – .39	0	1	1	1	1	2	2
Mean biserial	.64	.51	.53	.57	.55	.55	.54

Content categories	Number in: Form 0	Retained	New	Total
Equations	5	3	3	6
Geometry	4	2	3	5
Properties	7	4	4	8
Factoring	5	3	3	6
Arithmetic Reasoning	4	0	0	0

Table C-7. Difficulty Range, Biserial Range, and Content Categories for Mechanical Comprehension

Difficulty range	Officer[a] data Form 0			Officer difficulty estimated from Airmen data					
	Total		Retained	P1 New	P2 New	P1 Total		P2 Total	
	N	%	N	N	N	N	%	N	%
.80 - .89	1	5	0	0	0	0	0	0	0
.70 - .79	2	11	1	0	0	1	5	1	5
.60 - .69	3	16	3	1	1	4	20	4	20
.50 - .59	1	5	1	1	2	2	10	3	15
.40 - .49	10	52	4	7	5	11	55	9	45
.30 - .39	2	11	2	0	1	2	10	3	15
Mean difficulty	.51		.50	.48	.48	.50		.49	

Biserial range	Officer[a] total	Airmen		P1 New	P2 New	P1 Total	P2 Total
		Common	Retained				
.60 - .69	6	2	0	0	1	0	1
.50 - .59	7	4	2	1	1	3	3
.40 - .49	1	2	1	7	3	8	4
.30 - .39	4	5	4	1	4	5	8
.20 - .29	1	6	4	0	0	4	4
Mean biserial	.51	.37	.34	.45	.44	.39	.39

Content categories	Number in: Form 0	Retained	P1 New	P2 New	P1 Total	P2 Total
Illustrated						
Transfer of rotational motion	8	7	3	3	10	10
Rotational range of motion	6	3	2	2	5	5
Non-Illustrated						
Physics and physics terminology	1	0	0	1	0	1
Hardware and tools	2	0	1	2	1	2
Cars	2	1	3	1	4	2

[a] Officer Form 0 item statistics here are based on test results after deletion of one MC item.

Table C-8. Difficulty Range and Biserial Range for Electrical Maze

Difficulty range	Officer data Form 0 Total N	Officer data Form 0 Total %	Officer data Form 0 Retained N	Officer difficulty estimated from Airmen data P1 New N	Officer difficulty estimated from Airmen data P2 New N	Officer difficulty estimated from Airmen data P1 Total N	Officer difficulty estimated from Airmen data P1 Total %	Officer difficulty estimated from Airmen data P2 Total N	Officer difficulty estimated from Airmen data P2 Total %
.60 - .69	3	15	2	1	1	3	15	3	15
.50 - .59	4	20	1	3	3	4	20	4	20
.40 - .49	3	15	2	1	1	3	15	3	15
.30 - .39	1	5	1	0	0	1	5	1	5
.20 - .29	6	30	2	4	4	6	30	6	30
.10 - .19	3	15	2	1	1	3	15	3	15
Mean difficulty	.38		.38	.38	.38	.38		.38	

Biserial range	Officer total	Airmen Common	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.80 - .89	2	0	0	1	1	1	1
.70 - .79	5	6	3	4	4	7	7
.60 - .69	4	10	5	4	3	9	8
.50 - .59	6	3	2	1	2	3	4
.40 - .49	2	1	0	0	0	0	0
.30 - .39	1	0	0	0	0	0	0
Mean biserial	.63	.66	.65	.70	.69	.68	.67

There were no content categories for Electrical Maze.

Table C-9. Difficulty Range, Biserial Range, and Content Categories for Scale Reading

| Difficulty range | Officer data Form 0 | | | Officer difficulty estimated from Airmen data | | | | | |
| | Total | | Retained | P1 New | P2 New | P1 Total | | P2 Total | |
	N	%	N	N	N	N	%	N	%
.90 - .99	1	3	0	1	1	1	3	1	3
.80 - .89	3	8	3	3	3	6	15	6	15
.70 - .79	4	10	2	1	1	3	7	3	7
.60 - .69	5	13	3	4	4	7	18	7	18
.50 - .59	5	13	3	0	2	3	7	5	12
.40 - .49	8	20	5	5	2	10	25	7	18
.30 - .39	9	23	3	4	5	7	18	8	20
.20 - .29	4	10	1	2	2	3	7	3	7
Mean difficulty	.51		.56	.54	.54	.55		.55	

| Biserial range | Officer total | Airmen | | P1 New | P2 New | P1 Total | P2 Total |
		Common[a]	Retained				
.60 - .69	4	0	0	3	0	3	0
.50 - .59	20	1	1	6	4	7	5
.40 - .49	8	8	8	7	7	15	15
.30 - .39	6	6	6	4	6	10	12
.20 - .29	1	4	4	0	3	4	7
.10 - .19	0	1	1	0	0	1	1
Mean biserial	.49	.36	.36	.48	.40	.42	.38

Scale categories[b]	Number in: Form 0	Retained	New	Total
Straight	32	17	14	31
Curved	8	3	6	9
Equal	28	15	13	28
Log	12	5	7	12
Whole	27	12	15	27
Decimal	13	8	5	13

[a] The mean biserial shown for Form 0 common items was computed from experimental Sample 2 results only.

[b] Items are counted three times: once for the "straight or curved" dimension, once for "Equal or Log," and once for "Whole or Decimal."

40

Table C-10. Difficulty Range, Biserial Range, and Content Categories for Instrument Comprehension

Difficulty range	Officer data Form 0 Total N	Officer data Form 0 Total %	Officer data Form 0 Retained N	P1 New N	P2 New N	P1 Total N	P1 Total %	P2 Total N	P2 Total %
.60 – .69	2	10	0	2	2	2	10	2	10
.50 – .59	4	20	3	1	1	4	20	4	20
.40 – .49	7	35	2	5	5	7	35	7	35
.30 – .39	4	20	3	1	1	4	20	4	20
.20 – .29	3	15	2	1	1	3	15	3	15
Mean difficulty	.44		.42	.47	.46	.45		.44	

Biserial range	Officer total	Airmen Common	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.80 – .89	1	0	0	1	1	1	1
.70 – .79	4	2	1	0	5	1	6
.60 – .69	7	6	3	6	4	9	7
.50 – .59	7	8	4	3	0	7	5
.40 – .49	1	1	1	0	0	1	0
.30 – .39	0	2	0	0	0	0	0
.20 – .29	0	1	1	0	0	1	1
Mean biserial	.63	.54	.54	.63	.71	.59	.62

Content categories for Instrument Comprehension appear on the next page.

Table C-10. (concluded)

Special dimension categories	Number in: Form O	Retained	P1 P2 New		P1 P2 Total	
Dive/climb						
Dive	11	5	2	2	7	7
Level	5	2	4	4	6	6
Climb	4	3	4	4	7	7
Turn & Bank						
Left	8	5	3	1	8	6
Right	6	3	4	4	7	7
Forward	6	2	3	5	5	7
Direction						
N	1	0	3	3	3	3
S	4	2	1	1	3	3
E	8	4	3	4	7	8
W	7	4	3	2	7	6

Table C-11. Difficulty Range and Biserial Range for Block Counting

Difficulty range	Officer data Form O Total N	Officer data Form O Total %	Officer data Form O Retained N	Officer difficulty estimated from Airmen data P1 New N	Officer difficulty estimated from Airmen data P2 New N	Officer difficulty estimated from Airmen data P1 Total N	Officer difficulty estimated from Airmen data P1 Total %	Officer difficulty estimated from Airmen data P2 Total N	Officer difficulty estimated from Airmen data P2 Total %
.90 - .99	1	5	1	0	0	1	5	1	5
.80 - .89	2	10	1	0	0	1	5	1	5
.70 - .79	3	15	2	1	1	3	15	3	15
.60 - .69	2	10	1	2	1	3	15	2	10
.50 - .59	3	15	0	2	3	2	10	3	15
.40 - .49	1	5	0	2	2	2	10	2	10
.30 - .39	5	25	2	3	3	5	25	5	25
.20 - .29	3	15	3	0	0	3	15	3	15
Mean difficulty	.53		.52	.49	.50	.51		.51	

Biserial range	Officer total	Airmen Common[a]	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.90 - .99	0	1	0	0	2	2	0
.80 - .89	0	2	2	3	0	3	4
.70 - .79	10	5	4	3	2	7	6
.60 - .69	4	6	1	1	1	2	2
.50 - .59	4	3	2	1	3	5	3
.40 - .49	2	3	1	2	0	1	3
.30 - .39	0	0	0	0	2	0	2
Mean biserial	.65	.66	.68	.67	.62	.68	.65

There were no content categories for Block Counting.

[a] The mean biserial shown for Form O common items was computed from experimental Sample 1 results only.

Table C-12. Range of Item-Test Biserial Correlations for Table Reading[a]

Biserial	Speeded				Unspeeded		
range	Form 0[b]	P1	P2		Form 0	P1	P2
.90 - .99	1	2	2		1	6	8
.80 - .89	1	4	5		1	7	4
.70 - .79	1	3	2		3	2	5
.60 - .69	5	7	6		5	3	0
.50 - .59	4	4	4		4	2	2
.40 - .49	5	0	1		3	0	1
.30 - .39	3	0	0		3	0	0
Mean biserial	.56	.71	.70		.60	.81	.81

[a] Because of the speeded nature of the TR subtest, and its inherent lack of cognitive difficulty (items require locating values representing two coordinates), no difficulty estimates were made.

[b] All the Form 0 items selected as Common items were retained for Form P.

Table C-13. Difficulty Range, Biserial Range, and Content Categories for Aviation Information

Difficulty range	Officer data Form 0 Total N	%	Retained N	P1 New N	P2 New N	P1 Total N	%	P2 Total N	%
.70 - .79	2	10	1	1	1	2	10	2	10
.60 - .69	0	0	0	0	0	0	0	0	0
.50 - .59	4	20	2	2	2	4	20	4	20
.40 - .49	6	30	2	2	1	4	20	3	15
.30 - .39	6	30	3	2	4	5	25	7	35
.20 - .29	2	10	2	3	2	5	25	4	20
Mean difficulty	.44		.46	.40	.40	.43		.43	

Biserial range	Officer total	Cadet Common[a]	Retained	P1 New	P2 New	P1 Total	P2 Total
.70 - .79	3	5	0	1	1	0	0
.60 - .69	4	3	0	6	6	1	1
.50 - .59	9	5	0	2	2	2	3
.40 - .49	4	5	2	1	0	8	6
.30 - .39	0	1	5	0	1	6	7
.20 - .29	0	0	2	0	0	2	2
.10 - .19	0	1	1	0	0	1	1
Mean biserial	.56	.55	.55	.62	.61	.59	.58

Content categories	Number in: Form 0	Retained	New	Total
General	1	1	0	1
Navigation	1	0	1	1
Meteorology	1	1	0	1
Aerodynamics	2	1	2	3
Functions	3	2	1	3
Definitions	8	3	3	6
Hazards	1	0	1	1
Operations	3	2	2	4

[a] The mean biserial shown for Form 0 common items was computed from experimental Sample 1 results only.

Table C-14. Difficulty Range and Biserial Range for Rotated Blocks

| Difficulty range | Officer data Form O | | | Officer difficulty estimated from Airmen data | | | | | |
| | Total | | Retained | P1 New | P2 | P1 Total | | P2 | |
	N	%	N	N	N	N	%	N	%
.80 - .89	3	20	1	2	2	3	20	3	20
.70 - .79	0	0	0	0	0	0	0	0	0
.60 - .69	2	13	1	1	1	2	13	2	13
.50 - .59	1	7	1	0	0	1	7	1	7
.40 - .49	3	20	1	2	2	3	20	3	20
.30 - .39	3	20	2	2	2	4	27	4	27
.20 - .29	3	20	1	1	1	2	13	2	13
Mean difficulty	.50		.49	.52	.52	.51		.51	

| Biserial range | Officer total | Airmen | | P1 New | P2 | P1 Total | P2 |
		Common	Retained				
.70 - .79	0	1	0	0	0	0	0
.60 - .69	0	4	2	2	2	4	4
.50 - .59	3	4	2	4	5	6	7
.40 - .49	10	4	2	2	1	4	3
.30 - .39	2	2	1	0	0	1	1
Mean biserial	.46	.55	.53	.55	.54	.54	.54

There were no content categories for Rotated Blocks.

Table C-15. Difficulty Range, Biserial Range, and Content Categories for General Science

| Difficulty range | Officer data Form D | | | Officer difficulty estimated from Airmen data | | | | | |
| | Total | | Retained | P1 New | P2 New | P1 Total | | P2 | |
	N	%	N	N	N	N	%	N	%
.70 - .79	1	5	1	0	0	1	5	1	5
.60 - .69	1	5	0	1	1	1	5	1	5
.50 - .59	3	15	1	2	2	3	15	3	15
.40 - .49	7	35	4	3	3	7	35	7	35
.30 - .39	5	25	3	2	2	5	25	5	25
.20 - .29	2	10	1	2	2	3	15	3	15
.10 - .19	1	5	0	0	0	0	0	0	0
Mean difficulty	.42		.44	.43	.44	.44		.44	

| Biserial range | Officer total | Airmen | | P1 New | P2 New | P1 Total | P2 Total |
		Common	Retained				
.70 - .79	1	0	0	0	0	0	0
.60 - .69	4	0	0	2	0	2	0
.50 - .59	5	1	0	2	2	2	2
.40 - .49	5	5	5	5	4	10	9
.30 - .39	5	6	2	1	4	3	6
.20 - .29	0	7	3	0	0	3	3
.10 - .19	0	1	0	0	0	0	0
Mean biserial	.50	.32	.36	.49	.44	.43	.40

Content categories for General Science appear on the next page.

Table C-15. (concluded)

| Content categories | Number in: Form 0 | Retained | P1 | P2 | P1 | P2 |
			New		Total	
Astronomy	5	2	3	2	5	4
Biology	0	0	1	1	1	1
Chemistry	3	2	0	0	2	2
Geography	1	1	0	0	1	1
Basic Physics	1	0	1	1	1	1
Atomic Structure	2	1	1	1	2	2
Radiation	2	1	0	1	1	2
Electronics	0	0	1	1	1	1
Computers	0	0	1	1	1	1
Instrumentation	4	2	1	1	3	3
Measurement	1	0	1	1	1	1
Aviation Information	1	1	0	0	1	1

Table C-16. Difficulty Range, Biserial Range, and Content Categories for Hidden Figures

Difficulty range	Officer data Form 0 Total N	Officer data Form 0 Total %	Officer data Form 0 Retained N	Officer difficulty estimated from Airmen data P1 New N	Officer difficulty estimated from Airmen data P2 New N	Officer difficulty estimated from Airmen data P1 Total N	Officer difficulty estimated from Airmen data P1 Total %	Officer difficulty estimated from Airmen data P2 Total N	Officer difficulty estimated from Airmen data P2 Total %
.90 - .99	1	7	0	1	1	1	7	1	7
.80 - .89	4	27	2	1	0	3	20	2	13
.70 - .79	2	13	1	2	3	3	20	4	27
.60 - .69	0	0	0	0	0	0	0	0	0
.50 - .59	2	13	2	0	0	2	13	2	13
.40 - .49	3	20	2	2	2	4	27	4	27
.30 - .39	3	20	1	1	1	2	13	2	13
Mean difficulty	.63		.60	.66	.66	.63		.63	

Biserial range	Officer total	Airmen Common	Airmen Retained	P1 New	P2 New	P1 Total	P2 Total
.80 - .89	0	0	0	2	2	2	2
.70 - .79	0	3	2	3	3	5	5
.60 - .69	10	7	3	0	1	3	4
.50 - .59	5	5	3	1	0	4	3
.40 - .49	0	0	0	1	1	1	1
Mean biserial	.61	.63	.63	.71	.73	.67	.68

Content categories[a]	Number in: Form 0	Retained	New	Total
Figure 1	3	2	1	3
Figure 2	3	1	2	3
Figure 3	4	3	0	3
Figure 4	3	1	2	3
Figure 5	2	1	2	3

[a] Each content category is a distinct figure. A test item contains one of the figures embedded in a complex drawing.

49